GenAI-ML HW2
Build a Basic RAG System

TA: 馮柏翰, 林堅壬, 李梁玉軒
ntu-gen-ai-ml-2025-fall-ta@googlegroups.com

Deadline: 2025/**10/17** 23:59:59 (UTC+8)

Outline

- Task Description
- Dataset
- Evaluation Metric
- Baselines
  - Simple
  - Medium
  - Strong
  - Boss
- Submission & Deadline
- Grading Release Date
- Regulations

Links

- GENAI-ML 2025 FALL
- NTU COOL
- JudgeBoi
- JudgeBoi Guide
- Colab Sample Code
- Kaggle Sample Code
- **ML2025 Colab and Kaggle Tutorial**

Task Description - RAG as a Story

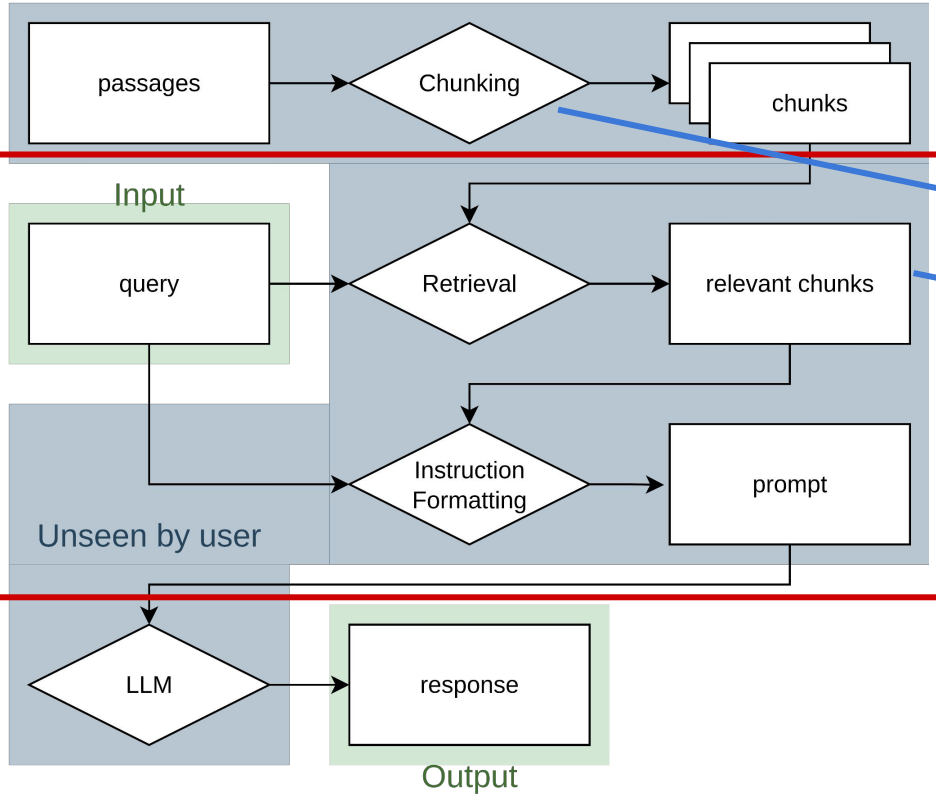
書櫃：存放書籍的知識庫

圖書館員：檢索與問題相關的書籍

老師：統整相關書籍回答問題

Figures are drawn by GPT

Task Description - RAG Workflow



1. Data Preprocessing

2. Retrieval

3. Generation

Rhombus: Function

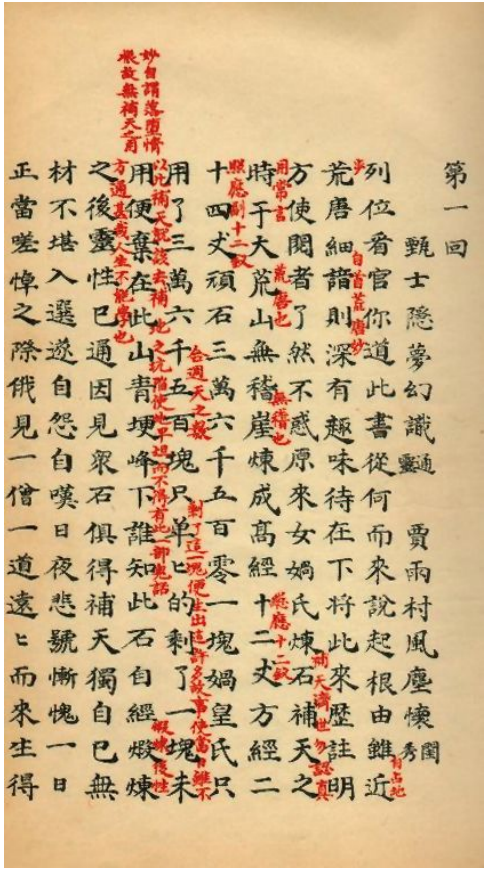
Rectangle: Text

Task Description - Introduction & Motivations

- In HW2, we need to build a basic Retrieval Augmented Generation (RAG) system
- What is RAG?
  - Given a query, **retrieve relevant documents(chunks)** from a knowledge base
  - **Combine relevant documents with generative models(LLM)** to produce coherent responses
- Why RAG?
  - **Reduce hallucinations** by grounding outputs in external, verifiable data
  - Handle domain-specific or up-to-date knowledge **without retraining**
  - **Protect private data** instead of embedding them into model's parameter
  - Give relevant information only, **saving time and computational cost** for the model to understand the entire knowledge base

Dataset - Metadata Information

- Knowledge base: Full text of《紅樓夢》
  - Source: INTERNET ARCHIVE
  - There are 120 chapters
  - Over 700,000 words
- There are 100 **(query, golden passage index, answer)** data pairs, each one is generated from 1 chapter using ChatGPT then checked by TAs
  - 50 data pairs in public test dataset
  - 50 data pairs in private test dataset


第一回
甄士隱夢幻識通
賈雨村風塵懷秀閨
列位看官你道此書從何而來說起根由雖近
荒唐細譜則深有趣味待在下將此來歷註明
方使閱者了然不惑原來女媧氏煉石補天之
時于大荒山無稽崖煉成高經十二丈方經二
十四丈頑石三萬六千五百零一塊媧皇氏只
用了三萬六千五百塊只單單的剩了一塊未
用便棄在此山青埂峰下誰知此石自經煅煉
之後靈性已通因見衆石俱得補天獨自己無
材不堪入選遂自怨自嘆日夜悲號慚愧一日
正當嗟悼之際俄見一僧一道遠ヒ而來生得

https://zh.wikipedia.org/zh-tw/%E7%BA%A2%E6%A5%BC%E6%A2%A6

Dataset - Public/Private Dataset

Imagine you have a double-sided exam paper, each side has 50 queries

- Public data pair example: `{"query": "賈政是賈寶玉的什麼親屬？", "golden_passage_index": 2, "answer": "父親"}`
- Private data pair example: `{"query": "賈寶玉最珍視、掛在身上的隨身之物是什麼？ "}`

	Before Deadline	**After Deadline**
Public Dataset	• For each data pair, (query, golden passage index, answer) is fully accessible • You can check your evaluation result whenever you want if you submit the prediction file to JudgeBoi	
Private Dataset	Only query for each data pair is released without golden passage index or answer	Evaluation result is released

Evaluation Metric - Final Score Calculation

- TAs will **cap your responses to 512 characters long**, then use GPT-based model to evaluate the quality of your responses in the prediction file
- Four types of scores for each response:

Score	Explanation
+1	**Fully correct:** same or almost the same to the answer
+0.5	**Partially correct:** some information is missing or wrong
0	**I don't know:** model says it cannot find the answer
-1	**Wrong:** response only contains hallucinated information

- Average of scores from all responses in public and private dataset would be your two final scores for grading your HW2

Evaluation Metric - Grading

There are **10 points** in total for HW2:

- **2 points**: Any successful submission to JudgeBoi
- **4 points**: Based on your final score on public dataset, 1 point for each public baseline
- **4 points**: Based on your final score on private dataset, 1 point for each private baseline

Evaluation Metric - Baseline Scores

- You can calculate your HW2 grade with following 8 baseline scores:

	Public	**Private**
Simple	-0.65	-0.60
Medium	0.20	0.00
Strong	0.30	0.25
Boss	0.40	0.35

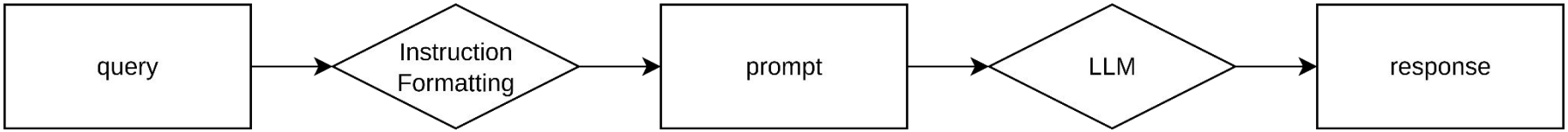
- For example, if your public/private final scores are 0.43/0.27, your HW2 grade will be:
  2(Any successful submission) + 4(Public Boss) + 3(Private Strong) = 9
- You can surpass all baselines within 2hr program execution time.

Baselines

- The following slides introduce methods to surpass corresponding baselines, while you are welcomed to modify every part of the sample code except:
  - **You can only use the LLM(unsloth/gemma-3-4b-it) provided by TAs for generation**
  - **You can only use open-sourced models to finish HW2. Any proprietary models or services that require API keys are not allowed**
- Disclaimer: Due to randomness, **you are not 100% guaranteed to surpass the baselines** if you try the methods mentioned by the following slides

Simple Baseline - Generation with LLM Only

- Just like how you chat with GPT/Gemini/Claude..., but without search tools
- Response quality only depends on internal knowledge of LLM



Medium Baseline - Sparse Retrieval + LLM

- Same as RAG Workflow using **BM25** as Retrieval function
- BM25 tokenize query and passages into **sequence of discrete tokens(sparse retrieval)** before similarity calculation
  - e.g., 資訊檢索系統可以從文件集合找到相關資料 → [資訊, 檢索, 系統, 可以, 從, 文件, 集合, 找到, 相關, 資料]
- Similarity is based on **keyword matching**
  - [**資訊**, **檢索**], [**資訊**, **檢索**, 系統, 可以, 從, 文件, 集合, 找到, 相關, 資料] → **higher similarity** 👍
  - [資訊, **檢索**], [這, 篇, 論文, 探討, 機器, 學習, 在, 影像, **檢索**, 領域] → **lower similarity** 👎

query chunks

- Pros
  - Works well without heavy domain-specific tuning
  - Formula is relatively simple and computationally efficient
- Cons
  - No Handling of Synonyms or Morphology by default (e.g, 地震 vs. 地牛翻身)
  - Cannot catch semantic relevancy (e.g., 只看字面不懂語意 vs. 詞不同時找不到相關內容)

Medium Baseline - BM25 Similarity

- Complete equation:

$$\text{score}(Q, D) = \sum_{t \in Q} \text{IDF}(t) \, \frac{f(t, D)\,(k_1 + 1)}{f(t, D) + k_1 \left(1 - b + b \, \frac{|D|}{\text{avgdl}}\right)}$$

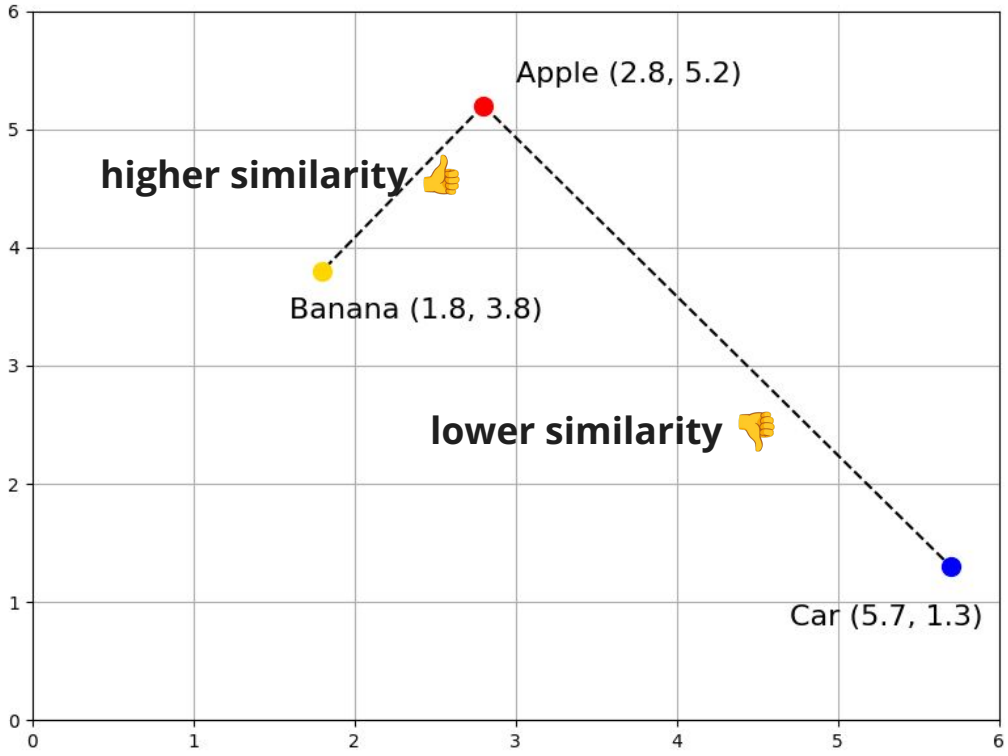
- Explanation from GPT5
- Source code implemented by Okapi
- Paper: The Probabilistic Relevance Framework: BM25 and Beyond

Strong Baseline - Hybrid Retrieval + LLM

- Besides BM25, add another Retrieval function using **embedding model**
- **Embedding model** tokenize query and passages into **sequence of float numbers(dense retrieval)** before similarity calculation
  - e.g., 資訊檢索系統可以從文件集合找到相關資料 → [5.6, 0.8, -7.2, ...] (**embedding**)
- Pros
  - Embeddings capture meaning at the phrase/sentence level, not just word frequency
  - Some embedding models work across languages or modalities (text–image)
- Cons
  - Training and inference require GPU
  - Out-of-domain performance may degrade without fine-tuning
- To leverage both BM25 and embedding model, we need to find a method to merge two lists of relevant chunks. (sparse + dense = hybrid)

Strong Baseline - Visualize Embedding Distance

- Similarity is based on **embedding distance**
  - Embeddings for objects
    - Apple: [2.8, 5.2]
    - Banana: [1.8, 3.8]
    - Car: [5.7, 1.3]
  - The closer the embeddings of two objects are, the higher similarity they get
- There are several ways to measure distance, and CosineSimilarity is used in HW2



Strong Baseline - Reciprocal Rank Fusion

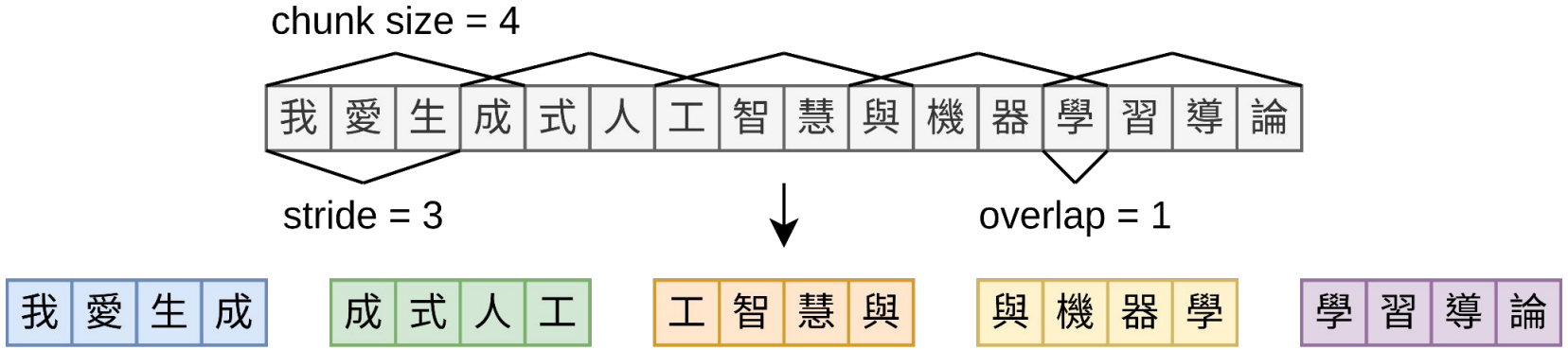
- Assume there are two ranked list of relevant chunks by BM25 and embedding model.
  - BM25: [c_1, c_4, c_3, c_5, c_2]
  - Embedding model: [c_5, c_1, c_3, c_4, c_2]
- Rank score of c_n is derived via the formula:
  $RRF(c) = \Sigma_{(r \in R)}\ 1 / (k + rank(c))$
  - **c**: chunk
  - **R**: set of retrieval models
  - **k**: constant (typically 60)
  - **rank(c)**: the rank of chunk **c**
- Final ranking: [c_1, c_5, c_4, c_3, c_2]

- $RRF(c_1) = (1 / (60+1)) + (1 / (60+2)) = 0.032522$
- $RRF(c_2) = (1 / (60+5)) + (1 / (60+5)) = 0.030769$
- $RRF(c_3) = (1 / (60+3)) + (1 / (60+3)) = 0.031746$
- $RRF(c_4) = (1 / (60+2)) + (1 / (60+4)) = 0.031754$
- $RRF(c_5) = (1 / (60+4)) + (1 / (60+1)) = 0.032018$

mathematical-intuition-behind-reciprocal-rank-fusion-rrf-explained-in-2-mins

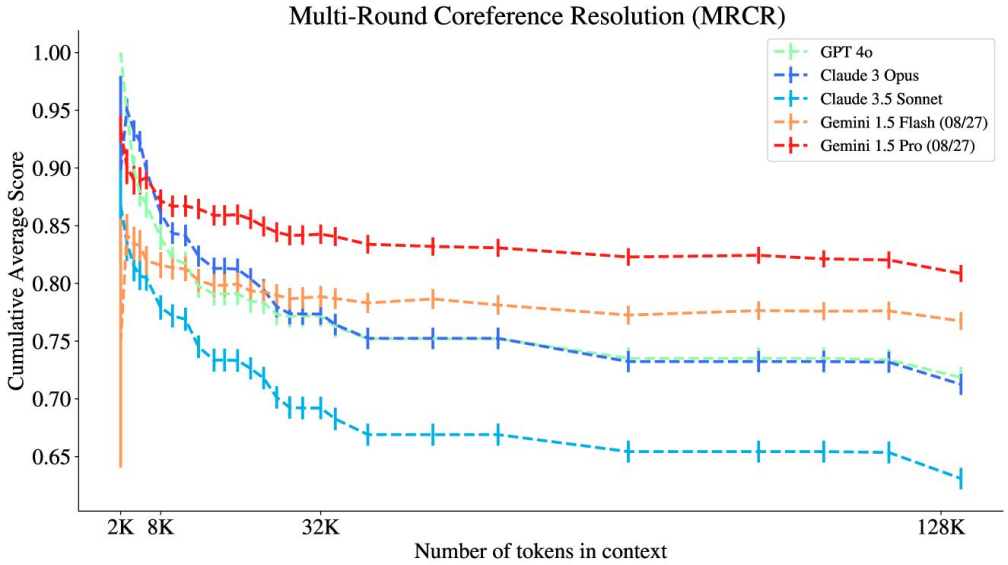
Boss Baseline - Better Chunking Strategy

- Adjust **chunk size** and **stride**
- **Chunk size**: Fixed length of each window processed at a time
  - Larger chunk size: More context per chunk for questions needing broader passages
  - Smaller chunk size: Tighter, more focused chunks for higher precision and relevance
- **Stride**: The step size between the start positions of consecutive chunks
  - Larger stride (less overlap): Faster indexing/querying and lower storage
  - Smaller stride (more overlap): Captures cross-boundary facts



Boss Baseline - Context Window

- Context window decides the largest number of token a LLM can handle
- LLM gemma-3-4b-it has 128k context window, but it does not mean that putting as many chunks as you can in the prompt within the context window is your best bet
- Recommend steps:
    1. Start with a small input size
    2. Gradually increase input size as performance is increasing
    3. Stop when you see a drop in performance



Michelangelo: Long Context Evaluations Beyond Haystacks via Latent Structure Queries

Submission & Deadline

- Submit your prediction file to **JudgeBoi**
- You have **3 submission quota per day**(evaluation is costly), reset at 12:00AM
- Evaluation time per prediction file is about 4 minutes if there are not too many students submitting at the same time
- Deadline: 2025/**10/17** 23:59:59 (UTC+8)
- **No late submission is allowed**, please finish you homework as soon as possible

Grading Release Date

- The grading of the homework will be released before 2025/11/07 23:59:59 (UTC+8)

Regulations

- You should NOT plagiarize
- You should NOT modify your prediction files manually
- Do NOT share codes or prediction files with any living creatures
- Do NOT use any approaches to submit your results more than 3 times a day
- Please protect your own work and ensure that your answers are not accessible to others. If your work is found to have been copied by others, you will be subject to the same penalties
- Your final grade x 0.9 and get a score 0 for that homework if you violate any of the above rules first time (within a semester)
- Your will get F for the final grade if you violate any of the above rules multiple times (within a semester)
- Prof. Lee & TAs preserve the rights to change the rules & grades

If You Have Any Questions

- NTU Cool **HW2** 作業討論區
  - 如果同學的問題不涉及作業答案或隱私，請**一律使用** NTU Cool 討論區
  - 助教們會優先回答NTU Cool討論區上的問題
- Email: ntu-gen-ai-ml-2025-fall-ta@googlegroups.com
  - Title should start with [GenAI-ML 2025 Fall **HW2**]
  - Email with the wrong title will be moved to trash automatically
- TA Hours
  - Time:
    - **9/22, 9/29, 10/6, 10/13 Monday 20:00~22:00**
    - **9/26, 10/3, 10/10, 10/17 Friday 17:30~19:30**
  - Location: **Google Meet**